

GENERALIZED HISTOGRAM INTERSECTION KERNEL FOR IMAGE RECOGNITION

Sabri Boughorbel[†], Jean-Philippe Tarel[‡], Nozha Boujmaa[†]

[†] INRIA Rocquencourt, IMEDIA, 78153 Le Chesnay, France

[‡] LCPC, DESE, 58 Bd Lefebvre, 75015 Paris, France

ABSTRACT

Histogram Intersection (HI) kernel has been recently introduced for image recognition tasks. The HI kernel is proved to be positive definite and thus can be used in Support Vector Machine (SVM) based recognition. Experimentally, it also leads to good recognition performances. However, its derivation applies only for binary strings such as color histograms computed on equally sized images. In this paper, we propose a new kernel, which we named Generalized Histogram Intersection (GHI) kernel, since it applies in a much larger variety of contexts. First, an original derivation of the positive definiteness of the GHI kernel is proposed in the general case. As a consequence, vectors of real values can be used, and the images no longer need to have the same size. Second, a hyper-parameter is added, compared to the HI kernel, which allows us to better tune the kernel model to particular databases. We present experiments which prove that the GHI kernel outperforms the simple HI kernel in a simple recognition task. Comparisons with other well-known kernels are also provided.

1. INTRODUCTION

The development of digital imaging systems remains mainly devoted to image acquisition and transmission. Another important issue is to help in organizing image databases for the users to be able to easily retrieve the desired images from a large amount of information. For this purpose, image recognition stands as a key discipline, since it provides automatic methods able to classify images using prior knowledge which we assume learned on a so-called training set. We only consider the two-class recognition problem, since multi-class problem can be viewed as multiple two-class problems. The images of the same object form the positive class, while the others form the negative class. The recognition process consists in two stages:

- Training stage: training samples containing labeled positive and negative images are used to learn algorithm parameters. Each image is represented by a vector $\mathbf{x}_i \in \mathcal{X}$ with label $y_i = \pm 1$, $1 \leq i \leq \ell$. ℓ is the number of samples. A boundary in \mathcal{X} is thus learned that separates positive from negative samples.
- Testing stage: the resulting classifier is applied to unlabeled images to decide whether they belong to the positive or the negative class.

SVM is proved to be an efficient learning algorithm for such a recognition task [1] since it leads to good generalization capacities [2]. With SVM, the training stage consists in minimizing the following quadratic problem with respect to parameters α_i :

$$W(\alpha) = - \sum_{i=1}^{\ell} \alpha_i + \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (1)$$

under the constraint $\sum_{i=1}^{\ell} y_i \alpha_i = 0$, where $K(\mathbf{x}, \mathbf{x}')$ is a kernel. Usually, the above optimization leads to sparse non-zero parameters α_i . Training samples with non-zero α_i are the so-called support vectors. Using the scalar product as kernel leads to linear discrimination. Using other kernels allows to take into account the non-linearity of the boundary in \mathcal{X} by performing an implicit mapping of \mathcal{X} towards a space of higher dimension. This is the so-called kernel trick [3]. The crucial condition that a kernel should satisfy to be suitable for SVM is to be definite positive, meaning that the SVM problem is convex, and hence that the solution of (1) is unique. Let us now recall the definition of positive definite kernels:

Definition 1. Let \mathcal{X} be a nonempty set. A function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a *positive definite* kernel if and only if it is symmetric (i.e $K(\mathbf{x}', \mathbf{x}) = K(\mathbf{x}, \mathbf{x}')$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$) and if:

$$\sum_{j,k=1}^n c_j c_k K(\mathbf{x}_j, \mathbf{x}_k) \geq 0$$

for all $n \in \mathbb{N}^*$, $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subseteq \mathcal{X}$, $\{c_1, \dots, c_n\} \subseteq \mathbb{R}$.

This work is partially supported by Muscle NoE.

General purpose kernels such as RBF kernel are widely used with success for different applications. Here, we introduce a new GHI kernel which have interesting properties, and that allows us to improve recognition results in our experiments with image histograms.

The paper is organized as follows. In Sec. 2, we recall the histogram intersection kernel and how it was proved to be suitable for SVM. We show the drawbacks and limits of this derivation. In Sec. 3, we introduce the generalized histogram intersection kernel and theoretical tools are presented and used to prove its positive definiteness. Then, we describe its advantages and interesting properties, compared with HI and other kernels. Finally, in Sec. 4, experimental results are presented comparing well-known kernels, HI and GHI kernels.

2. HISTOGRAM INTERSECTION KERNEL

Histogram Intersection (HI) has been first introduced in computer vision by [4]:

$$K_{HI}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^m \min\{x_i, x'_i\} \quad (2)$$

where $\mathbf{x} = (x_1, \dots, x_m)$ and $\mathbf{x}' = (x'_1, \dots, x'_m)$ are two color histograms with m -bins (in \mathbb{R}^m). It was successfully used as a similarity measure for image retrieval task. HI in (2) amounts counting the common number of pixels of same color between color histograms \mathbf{x} and \mathbf{x}' . Then, the sum over all colors is normalized by the size of the image. Experiments in [4] have shown that HI is robust to many transformations (distractions in the background of the object, view point changing, occlusion, for instance).

In [5], HI (2) is used as a kernel for SVM based recognition tasks. Let us review how HI kernel was proved to be positive definite. At first, we recall that any positive definite kernel can be rewritten as a scalar product. Indeed, a mapping function Φ can be always found such that $K(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}')$. In practice, such a mapping function do not need to be explicitly known. However, if a mapping function can be exhibited, the kernel is proved to be positive definite. This approach was followed in [5] to prove that HI kernel is positive definite. Within this approach, the proof is derived under two strong assumptions: 1) the used features are necessarily histograms, 2) all images must have the same size.

Following [5], let N denotes the number of image pixels. As all images have the same number of pixels, each bin of the histogram counts the number of pixels of corresponding color, and bin values are integer numbers. Therefore

they can be represented by the following binary code:

$$\Phi_{HI}(\mathbf{x}) = \left(\overbrace{1, 1, \dots, 1}^{x_1}, \underbrace{0, \dots, 0}_{N-x_1}, \right. \\ \left. \overbrace{1, 1, \dots, 1}^{x_2}, \underbrace{0, \dots, 0}_{N-x_2}, \dots, \overbrace{1, 1, \dots, 1}^{x_m}, \underbrace{0, \dots, 0}_{N-x_m} \right)$$

Each bin value x_i is encoded with N bits, such that the x_i first bits are equals to one and $N - x_i$ last bits are equal to zero. If we consider the dot product on this binary representation between $\Phi_{HI}(\mathbf{x})$ and $\Phi_{HI}(\mathbf{x}')$, we obtain for each bin than $\min\{x_i, y_i\}$ component products equal to 1 and $N - \min\{x_i, y_i\}$ equal to zero. Therefore we can write:

$$K_{HI}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^n \min\{x_i, y_i\} = \Phi_{HI}(\mathbf{x}) \cdot \Phi_{HI}(\mathbf{x}')$$

In [5], an extension to image with different sizes is also proposed by fixing the binary code length and normalizing histograms with respect to the image size. This extension is not correct since we cannot convert color histograms to integer value for all images. We thus propose in the next section a new proof of the positive definiteness of the HI kernel which overcomes these drawbacks. This new proof allows us to introduce an extra mapping on \mathcal{X} which is of main importance to improve performances, as shown in Sec. 4.

3. GENERALIZED HISTOGRAM INTERSECTION KERNEL

A type of kernels called conditionally positive definite kernels is proved to be suitable for SVM [6, 3]. Many useful links exist between positive definite and conditionally positive definite kernels which gives ways for deriving new interesting kernels of both types. We illustrate this by deriving that the Generalized Intersection Histogram (GHI) kernel is positive definite. To begin, we recall the definition of conditionally positive definite kernels:

Definition 2. Let \mathcal{X} be a nonempty set. A kernel K is said to be *conditionally positive definite* if and only if it is symmetric and if:

$$\sum_{j,k=1}^n c_j c_k K(\mathbf{x}_j, \mathbf{x}_k) \geq 0$$

for all $n \in \mathbb{N}^*$, $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subseteq \mathcal{X}$, $\{c_1, \dots, c_n\} \subseteq \mathbb{R}$ with $\sum_{j=1}^n c_j = 0$.

The only difference compared with positive definite kernel definition is that a supplementary constraint on c_j is added. A simple example of conditionally positive definite kernel is presented in the following proposition. It is of importance for the derivation of the GHI kernel.

Proposition 1.

$$K_f(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}) + f(\mathbf{x}'), (\mathbf{x}, \mathbf{x}') \in \mathcal{X} \times \mathcal{X}$$

is a conditionally positive definite kernel for any function f .

Proof. It is obvious that K_f is symmetric. Let us consider $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X}$, and $c_1, \dots, c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i = 0$, thus

$$\begin{aligned} \sum_{j,k=1}^n c_j c_k (f(\mathbf{x}_j) + f(\mathbf{x}_k)) &= \sum_{k=1}^n \overbrace{c_k}^0 \sum_{j=1}^n c_j f(\mathbf{x}_j) + \\ &\quad \sum_{j=1}^n \underbrace{c_j}_0 \sum_{k=1}^n c_k f(\mathbf{x}_k) = 0 \end{aligned}$$

Therefore, K_f is conditionally positive definite. \square

We now recall from [6], page 74, a theorem which links by a necessary and sufficient condition positive definite and conditionally positive definite kernels:

Theorem 1. Let \mathcal{X} be a nonempty set and let $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a symmetric kernel. Then K is conditionally positive definite if and only if $\exp(uK)$ is positive definite for all $u > 0$.

We now make use of this theorem to demonstrate that we can deduce another family of positive definite kernel that derives from conditionally positive kernels.

Proposition 2. If K is negative valued and a conditionally positive definite kernel then $\frac{1}{(-K)^\delta}$ is positive definite for all $\delta \geq 0$.

Proof. It is easy to prove the following:

$$\frac{1}{(-s)^\delta} = \frac{1}{\Gamma(\delta)} \int_0^{+\infty} u^{\delta-1} e^{su} du, s \leq 0, \delta > 0 \quad (3)$$

It follows from theorem 1 that if we consider a kernel K which is conditionally positive definite then e^{uK} is positive definite. As K is negative valued the integral in (3) is finite when we replace s by K . Therefore $\frac{1}{(-K)^\delta}$ is positive definite as a sum of positive definite kernels, for $\delta > 0$. For $\delta = 0$, the result is obvious. \square

Now, we have all tools that we need to derive the Generalized Histogram Intersection as a positive definite kernel. The following proposition gives our main theoretical result:

Proposition 3. The Generalized Intersection Histogram

$$K_{GHI}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^m \min\{|x_i|^\beta, |x'_i|^\beta\}, (\mathbf{x}, \mathbf{x}') \in \mathcal{X} \times \mathcal{X} \quad (4)$$

is positive definite for all $\beta \geq 0$.

Proof. Let us consider a positive valued function g . For any $\gamma \geq 0$, we construct the following function: $f(\mathbf{x}) = -g(\mathbf{x})^{-\gamma}$, $g(\mathbf{x}) \geq 0$, $\mathbf{x} \in \mathcal{X}$. It follows, from proposition 1, that the kernel $K_f(\mathbf{x}, \mathbf{x}') = -(g(\mathbf{x})^{-\gamma} + g(\mathbf{x}')^{-\gamma})$ is conditionally positive definite and negative valued. Thus, by applying proposition 2, we obtain that $K_{\gamma,\delta}(\mathbf{x}, \mathbf{x}') = (g(\mathbf{x})^{-\gamma} + g(\mathbf{x}')^{-\gamma})^{-\delta}$ is positive definite for any $\gamma, \delta \geq 0$. We now choose $\delta = \frac{1}{\gamma}$. From [6], page 69, the space of all positive definite kernels is close under point-wise limit. Therefore, when γ goes toward infinity, we obtain that:

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}') &= \lim_{\gamma \rightarrow +\infty} K_{\gamma, \frac{1}{\gamma}}(\mathbf{x}, \mathbf{x}') \\ &= \lim_{\gamma \rightarrow +\infty} \frac{1}{(g(\mathbf{x})^{-\gamma} + g(\mathbf{x}')^{-\gamma})^{\frac{1}{\gamma}}} \\ &= \frac{1}{\max\{g(\mathbf{x})^{-1}, g(\mathbf{x}')^{-1}\}} \\ &= \min\{g(\mathbf{x}), g(\mathbf{x}')\} \end{aligned}$$

is positive definite. By choosing $g(\mathbf{x}) = |x_i|^\beta$, we deduce that $K_{\beta,i}(\mathbf{x}, \mathbf{x}') = \min\{|x_i|^\beta, |x'_i|^\beta\}$ is positive definite. Finally, $K_{GHI}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^m K_{\beta,i}(\mathbf{x}, \mathbf{x}')$ is also positive definite as a sum of positive definite kernels. \square

Notice that contrary to [5], no restrictive assumption on image size or feature type are considered in the derivation of the GHI kernel. Compared to HI kernel, an hyper-parameter β is introduced in GHI kernel which adds an extra power mapping that can be used with advantages as shown in the experiments. The advantage of the power mapping was already observed in several contexts [1, 7]. In (4), absolute value are taken on every feature component. This means that GHI kernel is not the best for features of signed components. However, with the same proof, other interesting kernels can be derived with other choices of g .

Finally, it is important to notice that the GHI kernel leads to invariant SVM boundaries under scaling of all feature vectors. This property can be easily shown by extending [8] to the GHI and HI kernels as well. In practice, this means that when a factor λ is applied on all the samples $\{\mathbf{x}_i\}$, the boundary splitting positive and negative samples, obtained using the SVM, is the same up to the scale factor λ . This property is particularly useful, it is no more requested to select the scale of the kernel, which is the classical hyper-parameter of the RBF kernel.

4. EXPERIMENTS

On image recognition task, we compared performances of RBF kernel $K_{RBF}(\mathbf{x}, \mathbf{x}') = \exp(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2})$, Laplace kernel $K_{Laplace}(\mathbf{x}, \mathbf{x}') = \exp(-\frac{\|\mathbf{x}-\mathbf{x}'\|}{\sigma})$, GH kernel (2) and GHI kernel (4). The test is carried on an image database containing 5 classes from Corel database, with an additional



Fig. 1. Each row presents one of the 6 classes used for experiments (castles, sun rises, grasses, mountains, birds, water falls).

texture class of grasses, as shown in Fig. 1. Each class contains 100 images. Images are described using an RGB color histogram with a size of $4^3 = 64$ bins. A 3-fold cross validation is applied to estimate the errors rates. We considered the recognition problem of one class-vs-the others. We tuned two parameters to obtain the best validation error: 1) the SVM regularization coefficient $C \in \{0.0625, 0.5, 4, 32, 256, 2048\}$ and 2) the kernel hyper-parameter (σ for RBF and Laplace, β for GHI). Tab. 1 summarizes the average performances obtained for the four kernels.

Kernels	valid. error	test error
RBF	24.38 ± 0.54	24.33 ± 1.06
Laplace	23.50 ± 0.82	23.66 ± 0.89
HI	23.78 ± 0.08	24.00 ± 1.31
GHI	21.11 ± 1.29	20.33 ± 1.70

Table 1. Average validation and test errors for the four different kernels.

	castle	sun	grass	mount	bird	water
castle	69	5	0	12	5	5
sun	4	86	0	2	0	1
grass	0	0	100	0	1	0
mount	13	4	0	67	0	16
bird	5	2	0	5	86	7
water	9	3	0	14	8	71

Table 2. Best class confusion matrix using the GHI kernel.

We obtained that GHI kernel clearly improve performances compared to the RBF, Laplace, and HI kernels. As in [1, 7], the β optimal value is around 0.25. The improvement of the GHI is mainly due to the possible combination

of the 0.25 power mapping within the HI kernel. Finally, Tab. 2 presents the best class confusion obtained for the GHI kernel. Sunrises, Grasses and Birds classes are well recognized. Some confusions appear between Castles, Mountains and Waterfalls classes due to the presence of similar landscape colors.

5. CONCLUSION

We have proposed a generalization of the useful Histogram Intersection kernel. First, we have presented a review on how the Histogram Intersection kernel was previously proved to be positive definite and we have pointed two limitations of this derivation: images are considered having the same size and features are integer vectors. Second, we have proposed a generalization of the Histogram Intersection kernel, we named GHI kernel, by giving a very general and powerful proof of its positive definiteness. This demonstration overcomes the previously described drawbacks. Moreover, it allows us to introduced a mapping hyper-parameter which allows us to improve recognition performances in practice. One extra advantage of the GHI kernel is that it leads to invariant SVM boundaries under scaling of the feature space. Therefore, with the GHI kernel, it is unnecessary to optimize with respect to the scale hyper-parameter. Depending of the kind of used features, other kernels can be derived as positive definite from the proposed demonstration.

6. REFERENCES

- [1] O. Chapelle, P. Haffner, and V. Vapnik, "Svms for histogram-based image classification," *IEEE Transactions on Neural Networks*, 1999. special issue on Support Vectors, 1999.
- [2] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, 2nd edition, New York, 1999.
- [3] B. Scholkopf, "The kernel trick for distances," in *NIPS*, 2000, pp. 301–307.
- [4] M. Swain and D. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, 1991.
- [5] F. Odone, A. Barla, and A. Verri, "Building kernels from binary strings for image matching," *IEEE Transactions on Image Processing*, vol. 14, no. 2, pp. 169–180, 2005.
- [6] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*, Springer-Verlag, 1984.
- [7] J.-P. Tarel and S. Boughorbel, "On the choice of similarity measures for image retrieval by example," in *Proceedings of ACM MultiMedia Conference*, Juan Les Pins, France, 2002, pp. 446 – 455.
- [8] F. Fleuret and H. Sahbi, "Scale invariance of support vector machines based on the triangular kernel," in *Proceedings of IEEE Workshop on Statistical and Computational Theories of Vision*, Nice, France, 2003.